Feasibility study of applying descriptive ILP to large geographic databases

Joris Maervoet[1,2], Patrick De Causmaecker[2], Ann Nowé[3], and Greet Vanden Berghe[1]

[1] KaHo Sint-Lieven, Departement Industrieel Ingenieur, Vakgroep IT, Gebr. Desmetstraat 1, 9000 Gent, Belgium
[2] Katholieke Universiteit Leuven Campus Kortrijk, Faculty of Sciences, Department of Informatics, Etienne Sabbelaan 53, 8500 Kortrijk, Belgium
[3] Vrije Universiteit Brussel, Department of Computer Science, Computational Modeling Lab, Pleinlaan 2, 1050 Brussel, Belgium

Abstract. This paper discusses a case study in which the aim is to discover regularities and anomalies in large databases containing geographic data, to improve and maintain the overall data quality. The application of Inductive Logic Programming (ILP) and descriptive ILP in particular to this case is discussed and motivated. In an experiment on real-world data, a classical descriptive ILP algorithm (WARMR) is applied to the hamlet of Beggen, Luxemburg, to mine for rules describing regularities. The algorithm adopts the setting of learning from interpretations. In a next stage, the violating interpretations of the rules could be traced to identify candidate anomalies. A rule export module was set up to feed the results to a rule checking engine for further validation of this experiment. Finally, the results are discussed, the feasibilities of the system used in the case study are assessed and possibilities w.r.t. a larger scale application of the experiment are discussed.

1 Introduction

The Quality Gate project aims to develop a system that is able to maintain the quality of data located in the central database of the company Tele Atlas, which is a geographic data provider. For this purpose, the project focuses on data mining techniques to extract rules and regularities automatically out of the spatial data and on tracing according anomalies.

This work accounts for the feasibility of applying ILP and its subdomains to the geographic data in order to control the quality of data. ILP is a relational data mining technique in which input data as well as output patterns have a relational representation: the language of logic programs (or similar). This makes ILP an interesting candidate to apply to Quality Gate. A possible approach would be to employ supervised learning and look for explanations of given examples of irregularities. Another approach is to use unsupervised learning and look for regularities in the data in the form of rules, of which the violations

are then candidate irregularities. The latter approach offers the most interesting functionality and is the one used in our case study. It is a form of indirect unexpected knowledge mining. Plantevit et al.[1] classify methods for unexpected rule discovery into user-driven methods (intervention by a human expert) and data-driven approaches, which are again subdivided into using unexpectedness-oriented measures on classical algorithms and new algorithms integrating new concepts of unexpectedness. In the case study we investigate the data-driven approach using unexpectedness-oriented measures on classical algorithms.

Outlier detection in multidimensional data has been studied extensively in the domains of statistics and automated learning. A lot of the methods proposed are based on proximity [2] or density [3] analysis. Another approach is to cluster the data first and to identify the entities that lie outside any cluster as outliers. Aggarwal and Yu [4] proposed the use of evolutionary algorithms to discover outliers in high-dimensional input data. However, our geographical data needs a more relational approach because the input data consists of several interrelated data types. Therefore, approaches originating from the field of ILP, in which patterns are learnt from relational data, are more likely in these circumstances.

Koperski and Han [5] were the first to define the concept of spatial association rules and to introduce an appropriate mining method. Further related work in geographic rule mining has been carried out at the University of Bari, Italy. In this work, the geographic hierarchy is integrated in the mining process. In [6] and [7], census data from Stockport is analysed by learning rules about socio-economic issues (e.g. commuter habits) in small districts to support transport planners. Furthermore, Lisi and Malerba [8] designed a hybrid language, called AL-log, that allows relational and structural description of data in order to use ILP to induce multiple-level association rules. The system is more performant than classical ILP with hierarchical information as background knowledge because taxonomic reasoning is integrated directly in the search process. In [9] geographical concepts are learnt which are not explicitly modelled. The system provides end-users with a tool for the automatic recognition of morphological elements (e.g. a fluvial landscape). In [10], geographic impact factors are learnt for rent prize categories in Munich. Most of the work is supervised learning. Another difference with the approach of our case study is that eventually some spatial knowledge is put into the preconditions of our experiment (i.e. the interpretations) and that our outcome rules do not necessarily contain literals describing spatial relationships. Furthermore, our case has special interests in unexpected knowledge mining.

Another related approach has been explored by Kuramochi and Karypis [11], who applied finding frequent patterns on graph-modelled input data successfully. In this case the mining procedure is optimised using graph properties such that uninteresting subgraphs can be pruned, candidate patterns can be generated and frequencies can be counted efficiently. Again, it differs with our work

because we are not necessarily looking for rules with spatial relationships. Furthermore, ILP goes beyond the propositional rule learning employed by [11].

The next section presents the problem of quality control for geographical data and its main challenges. A rough sketch of the company's data model is given, together with some sample rules. Section 3 motivates the application of ILP and its subdomains to Quality Gate. The fourth section outlines an experiment in which the WARMR algorithm is applied to a modified subset of the database. It is shown how to look up anomalies and export outcome rules into the representation used by the company. Section 5 discusses the experiment of the previous chapter. Encountered problems are identified. It presents suggestions on how to apply the algorithm on a larger scale. The last section is the conclusion of this work and contains some suggestions for further research.

2 Problem description

2.1 Quality control for geographical data

The company Tele Atlas is market leader in the provision of geographic data to other companies. The data supports three types of applications: navigation systems, Geographic Information Systems (GIS) and applications that provide Location Based Services (LBS). In these application areas there is an increasing demand for data quality. In certain situations companies could even be legally liable for inaccuracies in geographic data.

The company uses several methods to collect geographic data. Beside the processing of satellite images and data originating from other organisations (such as the government), the distribution of mobile mapping vehicles is the most important way to collect data. These vehicles are controlled by mobile specialists who proceed to regions of interest to create, to validate or to update field data. The specialist is provided with a set of tools that support the semi-automatic acquisition of data.

Most of the anomalies that reside in the database originate from human mistakes or from inconsistency between different sources. Moreover, not all kinds of mistakes are immediately traceable. Some only become apparent when a considerable number of updates are combined. Others need an elaborate search in the central database.

In a recent manual test case, data engineers started to investigate the intrinsic logic behind the spatial data, based on anomalies they encountered in the past. This process yielded a substantial number of rules, despite the short period during which it has been carried out. So the number of rules makes rigid checking of DB updates (manually entered by the specialists) for complete consistency impossible. Furthermore it appeared that more general trends are much more difficult to formulate. There is a general belief that various heuristics could help

Fig. 1. Geographic data collection with mobile mapping vehicles

in finding inaccuracies (e.g. elaboration on same rule, checking the geographic neighbourhood, updates by the same actor). Currently, the company investigates the applicability of a rule engine in their software, which enables the automatic tracing of anomalies. However the construction of rules is a manual process, driven by knowledge about known anomalies. Due to the high number of concrete rules, the rule base is flat and hardly manageable.

2.2 Data model and rule examples

The problem data basically consists of complex geographical features e.g. a restaurant or a water area. These complex features map to simple features which are points, lines, areas or some combination of these. The simple features can be seen as a combination of nodes, edges and faces in the spatial plane. A complex feature also has non-spatial attributes, which are organised in a hierarchical manner (e.g. composite address of a restaurant). Furthermore, there is aggregation and multiple inheritance between feature types. Extra relationships (e.g. allowed traffic manoeuvre) between features can be imposed using associations. The most common feature type is the road element, which represents in fact a part of a road (not necessarily from a physical junction to another). It maps to a line feature and has functional road class, specifying road importance, amongst its attributes.

Some examples of regularity rules that were manually discovered by data engineers are listed below.

1. *At each roundabout, not all connected roads have the same single direction of traffic (all inwards or all outwards).* Violations to this rule occur when a traffic direction update has not been executed over all connected roads or after a typing error. In this paper it will further be referenced to as the graveyard rule.

2. *A road element that is at both sides connected with road elements that have name X, has name X as well.* This rule is violated when an operator executes an incomplete name update. However in some exceptional situations, such a discontinuity is realistic. Both this and the previous rule link same-type features lying closely together and compare same-type non-spatial attributes of them.
3. *Each highway road element is connected at both sides.* Violations to this rule occur for example in the intersection of the highway with the border between two regions that are operated by two different mobile specialists and are due to coordination problems. It links same-type features lying closely together.

2.3 Challenges

For this application, there are two major challenges with regard to the mining task of rule and/or anomaly discovery, regardless of the chosen approach.

First of all, the geographic database is gigantic in size, rich in structure and highly relational. So scalability is very important to apply the KDD process successfully. Provost and Kolluri [12] have made a survey of strategies for scaling up the kind of KDD process that employs inductive algorithms. However, most of the strategies stated can be easily generalised to other types of data mining and apply for our case. The large database size could be handled by various data partitioning strategies, with regard to data instances and/or data features. A rich data structure often yields a large pattern space. Strategic cropping of, as well as heuristic search through the pattern space could help. The relational data representation, which can be integrated in the KDD mining process, can be omitted by flattening the data or can be shifted towards the database management level.

Secondly, the spatial nature of the input data should be dealt with. Knowledge discovery in geographic databases is more difficult than traditional KDD. Shekhar et al. [13] even doubt the usefulness of traditional data mining techniques in this context because of data type complexity and intrinsic spatial relationships. According to [13], spatial data mining differs from classical data mining w.r.t. data input, statistical foundation, output patterns and computational process. Geographic data is composed of objects with spatial and non-spatial attributes. Because "everything is related to everything else but nearby things are more related than distant things", materialisation of spatial relationships will be necessary and will determine the usefulness of the results to a certain extent. However, finding irregularities in our geographic dataset is not restricted to a purely spatial data mining problem. Not all anomalies can be detected using a spatial framework (e.g. discovery of a wrong speed limit based on the road type).

3 Applicability of ILP to Quality Gate

3.1 Motivation

Inductive Logic Programming (ILP) is a relational data mining technique that integrates the relational representation of the input data in the mining process[14] and in which input data as well as output patterns have a relational representation: the language of logic programs (or similar). It involves the explicit construction of First-order Logic rules from inherent regularities and trends in the input data. ILP algorithms can be seen as a search in a space of hypotheses that describe or classify parts of the input data[15].

ILP naturally comes as a candidate technique to apply to Quality Gate because automatic rule induction embodies the process of discovering (ir)regularities and trends in the geographical data, using an explicit representation that allows to control and loop up new occurrences of anomalies in the data and offers opportunities to be integrated in the company's business process. Because the geographic data is highly relational, ILP is preferable over propositional learning systems.

3.2 ILP subdomains

Mainly, ILP systems could be categorised into systems for predictive and for descriptive induction [16].

Predictive data mining is described as a discipline in which a hypothesis is induced that correctly classifies some given positive and negative examples (observations). So the outcome hypotheses of predictive ILP will be used for classification and prediction. Of course, this representation of classifiers could be extended to classification into any finite number of classes. Some predictive ILP systems induce LP rules, others generate decision trees that can be translated to small logical programs using cuts [16] (or using negation as failure). Predictive data mining has the longest tradition. Historically, predictive ILP systems are divided into empirical and interactive ILP (respectively EILP and IILP) systems, which have different properties, as described in [15].

The aim of descriptive data mining is to find regularities within a given set of unclassified examples [16]. In particular, as many usable data characteristics as possible are collected to build a sort of most specific hypothesis that describes the entire set of examples. Note that predictive induction can be handled more efficiently (lower complexity) than descriptive induction.

3.3 Quality Gate in relation to ILP subdomains

Predictive ILP systems, and specifically EILP systems, are interesting to learn rules driven by known anomalies. Anomalies are negative examples, other data is

positive (or the other way around). The outcome hypotheses can be used to classify new data into the category of anomalies or correct data. However, there can be a problem with the significance of known anomalies compared to the number of unknown anomalies that are presented to the system as positive examples. It is not possible to learn rules about regularities using EILP without the guidance by known anomalies, because a generation of positive examples using the closed world assumption (CWA) does not work here.

EILP systems are commonly non-interactive 'from scratch' learners, which enable the assumed automatic rule generation. They need a large example set, which is no problem for our application. The fact that they are single predicate batch learners requires a higher-level strategy to support scalability.

At first glance, IILP systems are interesting candidate systems for Quality Gate because of their incremental behaviour. However they are inappropriate for the intended fully automatic generation of rules due to the fact that they rather revise than construct theories and that they count on external parties to judge or review newly generated examples or hypotheses. Nevertheless a nice support tool for domain experts who have low I.T. affinity can be developed using IILP systems. It provides the user with a GUI that shows some geographic situations that should be approved or disapproved or that should be commented. Like in a Mastermind game, the system is challenged to find the solution by asking as few questions as possible.

DILP systems are interesting to learn rules that describe intrinsic logic and trends in the geographic database. It is a form of unsupervised learning. The outcome consists of sets of all possible hypotheses that satisfy the constraint set. A much larger part of the hypothesis space is searched, because DILP is not guided by positive and negative examples and because of the result diversity. This makes DILP algorithms intrinsically less performant. However, DILP is the most suitable technique to apply to our problem, because there is no knowledge about anomalies needed in advance. Therefore, this technique is applied in the experiment discussed in the next sections.

4 Case study

4.1 Our approach based on WARMR

Stolle et al. [17] offer a formal generic definition on the descriptive ILP setting: the aim of DILP is to find a set of clauses $Th(cons, E, L_h)$ in a language L_h that hold over a set of interpretations E and satisfy a predefined conjunction of constraints $cons(h, E)$. The background knowledge B is used during hypothesis construction and embodies prior knowledge about the learning problem. A DILP algorithm typically uses one specific conjunction of constraints. A non-exclusive list of possible constraints is given in [17]. The definition adopts the partition of the input data set into interpretations to scale up the algorithm (as used in [16]

and [18]).

WARMR is a DILP algorithm that finds patterns satisfying the constraint query $freq(covers; E) > t$, as described by Stolle et al.[17], which means that more than t elements of the set of interpretations E should be covered by the pattern. WARMR involves a level-wise discovery of frequent datalog patterns, which are function-free conjunctive formulas [8]. These patterns are commonly post-processed into rules or query extensions. More details on the algorithm can be found in Dehaspe[19].

In this experiment, WARMR will be used to construct query extensions that describe regularities in the input data. The approach is outlined in the figure below. As an example, the figure shows a map of connected road elements of which one allowed traffic direction is anomalous and creates a so-called inverse graveyard in node a. In a next step, the map is partitioned into interpretations by grouping road elements around the nodes. The data is translated into a Prolog format, in which separate predicates are used to declare the existence of features and to define properties of these features. The hypothesis language is defined. Additionally, background knowledge, which is used during rule construction, can be defined. Next, the WARMR algorithm is run to construct frequent queries. During postprocessing, query extensions are constructed. In a next stage, these rules are presented to human users and can be accepted or rejected. Query extensions with high but not 100% confidence are of particular interest to trace according violations.

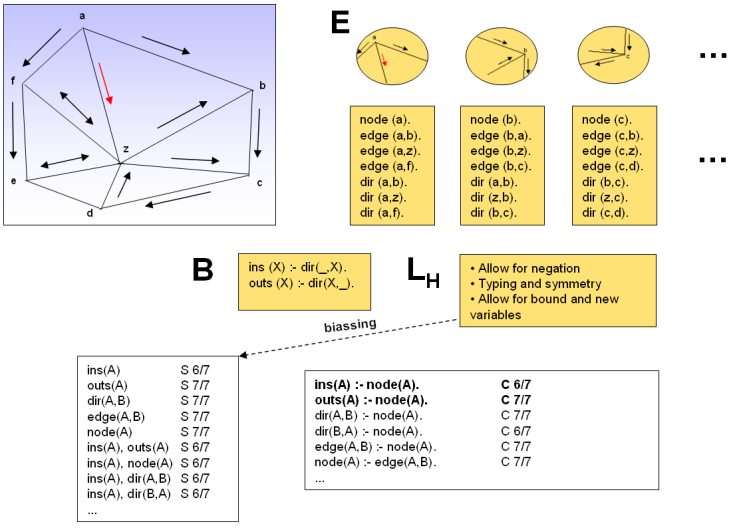

Fig. 2. Applying WARMR to geographic data: general approach

Notice that the coverage constraint employed by WARMR is existentially quantified within the interpretations. When for example interpretations symbolise separate villages containing roads, the rule *large(A) :- road(A)* does not imply that all roads in the villages are large, but that there are large roads in a village if there are roads.

The anomalies we mine can be defined as interpretations for which B is true and H is false given a rule H :- B with high confidence and support above a certain threshold. In other words, taking the WARMR coverage constraint into account, we are looking for interpretations e, given a clause of the form h :- b_1 , ... , b_m for which $\exists\theta : \{b_1\theta, ..., b_n\theta\} \subseteq e \Rightarrow \neg\exists\sigma : \{b_1\sigma, ..., b_n\sigma, h\sigma\} \subseteq e$ amongst a substantially larger group of interpretations e for which $\exists\theta : \{b_1\theta, ..., b_n\theta\} \subseteq e \Rightarrow \exists\sigma : \{b_1\sigma, ..., b_n\sigma, h\sigma\} \subseteq e$, in which θ and σ represent substitutions.

4.2 Data preparation and modelling: the Beggen data

In this experiment, we intend to induce rules that relate one or more geographic objects and mainly non-spatial attributes of these objects, optionally using the spatial relationship of immediate proximity. We restricted ourselves to junctions (incl. geometry) and road elements (incl. names, geometry and importance) in the village of Beggen, Luxemburg. For experimental reasons, another form of data construction is applied: the generation of random data. A small part of the junctions was randomly left out to show that the system works for imperfect data. Allowed traffic flows were randomly generated because the Beggen data set does not contain a significant number of single flows to induce the graveyard rule(s). The input data was partitioned into interpretations by taking the immediate neighbourhood of each original junction. This resulted in an input data set of 68 interpretations described in Prolog format, using a separate predicates to announce the existence (by ID) and geometry of junctions and road elements and to define properties. The background knowledge only consists of predicates that are negations of predicates used in the knowledge base.

The most important WARMR parameters are: minimum allowed support of frequent queries (0.05), minimum allowed confidence of resulting rules (0.7) and search depth (5 literals). Also the language bias is specified at this point. Type constraints are specified for the predicates described above and for their negated versions, specified in the background knowledge. Mode constraints allow for all predicate parameters to introduce new variables (+), to bind to already introduced variables (-) and to introduce new variables that are strictly different from a previously introduced same type variable (\). Functional road classes are allowed to be formulated as constants, and (not_)junction predicates can only appear once in the hypotheses because each interpretation contains at most one junction. Occurrence constraints ensure that road element flow predicates on the same variable only appear once in the hypotheses.

The WARMR run took about 15 minutes on a standard PC and, for the settings

given, 39 696 frequent queries were generated (maximum length 5), out of which 13 584 query extensions were generated.

4.3 Results

About 30 interesting rules were manually selected from the large set of query extensions. Three classes of interesting regularity rules (query extensions) are listed below, together with some example rules. The classification is rather intuitive.

1. **Rules on completeness.** A first group of example rules are rules that state a simple (co)existence of one type of objects and/or its properties in the interpretations e.g. *Interpretations containing a class 1 road description, contain a name definition for that road element (confidence: 100%; support: 30.8%).* Note that, when the rule concerns properties of the same feature, the use of one interpretation per feature would have been more appropriate. Other rules in this series are rather diagnostic e.g. *Interpretations containing a road element functional road class description but without name definition for the road element involved, contain a class 7 road element (confidence: 100%; support: 7.3%).*
2. **Rules on flows.** This type of rule links the existence of allowed traffic flows within the interpretations. The graveyard rule was (re)discovered in two parts, because in our system only one literal is allowed in the consequence of an implication: *Interpretations not containing any 'both direction' road elements contain a flow from centre road element (confidence: 96.5%; support: 41.1%).* and *Interpretations not containing any 'both direction' road elements contain a flow to centre road element (confidence: 89.6%; support: 38.2%).*
3. **Rules on interpretation statistics and continuity.** This type of rule relates the existence of distinct objects and/or properties within an interpretation. Some rules can only be used for statistics e.g. *Interpretations containing a road element, contain another road element with different ID (confidence: 80.8%; support: 80.8%).* So 80.8% of the interpretations contain more than one road element. Others refer to laws on the continuity of road properties through interpretations e.g. *Interpretations containing a class 1 and a class 4 description, contain a class 1 description for a strictly different road element (confidence: 100%; support: 5.8%).* Note that most of the violations against these rules could be avoided by the use of border predicates, depicting the geographic objects located at the border of Beggen.

The outcome of this experiment is a giant number of hypotheses, which is not manageable to present to a human reader nor suitable to use immediately for anomaly detection. In [20] some suggestions are listed to reduce the number of rules drastically either by specifying constraints in order to limit the number of rules generated or by means to filter the rules obtained.

4.4 Tracing anomalies and rule export

The case study has been extended by two modules to improve the usability and to enable further validation by experts.

A first module allows the end user to look up violating interpretations for each generated rule that has a confidence lower than 100%. This is realised by a Prolog interpreter, for which the background knowledge was asserted. If a given rule *H :- B* is looked up, for each interpretation I, sequentially, the facts of I are asserted; if the query B succeeds and the query B,H does not, the interpretation is added to the set of violating interpretations and the facts of I are retracted. The violations can be easily visualised in a spread sheet version of the Beggen map, because the interpretations were given coordinate names.

A second module is a rule export module. It translates the outcome rules into the XML format used by the company. During rule translation, the whole context in which the experiment was set up should be integrated in the outcome rules. For example, the original Prolog rule *roadelement_name(C,B), not(=(C,A)) :- true, roadelement_name(A,B)* is, making abstraction of the syntax and data model details, translated into

Check for each junction: if there exists a simpleorderarea8 containing the junction that has the officialname Beggen, and there exists a roadelement A that touches the junction and has an officialname, then there exists a roadelement C that touches the junction and has an officialname and the officialname of C equals the officialname of A and the ID of C is different from the ID of A.

A first component that is reflected in the global rule structure is the WARMR constraint conjunction. Furthermore, the geographic area (Beggen) for which the rule applies should be adopted. Also the way the interpretations were constructed has its impact on the main root predicate of the rule and on the relation that is declared between the root feature and new geographic features in the rule body. Furthermore, cardinalities in the data model influence the lower-level formulations (not in the example; e.g. "there exists a property vs. has as property"). Also implicit Prolog bindings should be taken into account and other Prolog elements such as negation and equality should be translated correctly.

5 Discussion

The ultimate aim of this research is to build an operational component that induces a confined set of readable rules that hold over the geographical data in a semi-automatic manner. It must be a generic component that mines the relations between given object and property types based on a given geographic region. The experiment showed us a partially successful application of a descriptive ILP algorithm. However, still a lot of problems need to be solved in order to build the intended generic component.

5.1 A reflection on the experiment

DILP goes clearly beyond intra-feature learning and is suitable to discover relationships among geographic features and attributes. The case study ended up with a rule overhead, for which we proposed suggestions to reduce the number of rules. The data representation of LP requires a serious preprocessing effort but enables an easy representation of outcome patterns and integration of background knowledge. Most of the available DILP systems require a considerable effort for the developer to get acquainted with and are rarely robust.

The resulting rules of the WARMR algorithm satisfy a constraint with existential quantification. This existential quantification yields a substantial performance improvement, but constrains the validity of the induced rules to the level of the partitioning chosen. For existential quantification, it seems preferable to omit all the rules that do not strictly correspond with the partitioning intentions. On the other hand, inducing rules with universal quantification is much less performant but the intended range of rules is much larger for one single experiment. But existential quantification requires more effort in the planning of the interpretation partitioning and in the formulation of hypothesis restrictions. In both situations results become more precise when a border predicate is used to indicate the interpretation's boundaries.

The use of negation as failure (not-literals) in the induced hypotheses caused some problems. Some algorithmic problems were encountered, because bound not-literals are order-dependent within the hypothesis. When introduced as a first part of the binding or unbound, the negated literal refers to existence within the interpretation; otherwise it refers to a failing property of an already introduced variable. In fact, for this research, both usages are interesting. Note that the use of multiple disjunct literals in the head yields the same logical expressiveness as the use of negated literals. Each disjunction in the head can be expressed as single literal headed rules using negation in the body. The use of not-literals provides more information, whereas the use of disjunctions in the head yields a smaller hypothesis space.

The rule export module contributed to the interpretability of the resulting rules for end users, who are domain specialists but not necessarily computer scientists. The next question considers whether sets of rules can be logically combined to improve the inventory and usability of rules.

5.2 Towards a larger scale application of the experiment

At this point it is clear that it is not possible to set up one big ILP experiment that mines for all types of rules, nor one that mines for all object and property types, nor one that takes a whole country as input. So a more general strategy in planning series of ILP experiments is required. Such strategy should mainly manage input space and pattern space scalability.

A first requirement is a more generic representation of objects and properties types within the DILP experiment to mine for any object and property types of choice. This could be done by the integration of the metadata in the predicate scheme construction during the data formatting task. Measures should be taken with regard to the language bias to avoid that meta-information is mined. In fact this is a first step to integrate the existing concept hierarchies more in the search process.

Scalability of the input space is a minor problem in the organisation of series of ILP experiments. Thanks to the learning from interpretations setting, WARMR's time complexity is linear w.r.t. the interpretation set size for an individual experiment. Some relational data mining systems also have random sampling of input data (interpretations) as an option integrated in the search process. Of course it is not possible to present large geographic regions to a single ILP system, but in many cases it may not be necessary to do so. The graveyard rule, for example, could have enough support in a large city. Maybe further checking in rural areas would not even increase its support, because of the lack of single traffic flow roads. Therefore, a possible solution is input data sampling during the data selection task. This sampling can be random, but in case of geographic data it might be better to select geographic regions for which certain conditions hold, such as variance or the presence of selected object types.

Pattern space scalability is the most delicate issue of a DILP application, because the size of the hypothesis space is exponential w.r.t. the number of allowed hypothesis predicates for an individual experiment. DILP systems such as ACE enable query sampling. It implies that not all possible frequent queries are generated and executed; so it results in missing rules but also in a large reduction in execution time. However each form of knowledge about probable and improbable object type, property type and advanced features combinations is welcome. In this manner a series of smaller DILP experiments can be set up so that the global size of pattern spaces decreases enormously. Probably the concept hierarchy could help to crop useless type combinations. Another option is to organise the type combinations in relation to real-world map formats: low-scale national map, regional tourist map, map for geologists. This increases the probability to discover realistic relationships between object and property types.

6 Conclusion

The purpose of this research is to study the applicability of (D)ILP to discover anomalies in large geographical databases, by means of a case study on a realistic data sample. ILP appeared to be an interesting mining method for knowledge discovery in geographical databases because it enables the direct induction of usable first-order logic rules and copes with the relational nature of the input data. When rule learning were driven by known anomalies, which is not the case

here, EILP systems would be applicable. In the case of unsupervised learning, DILP systems are used to describe intrinsic logic in the geographic database. In the latter case, it is not possible to build and consider a magna carta of regularities that hold over the input data, but it is worth to consider almost complete rules and look up their violations afterwards.

The experiment shows the intrinsic value of DILP. However, additional fine-tuning and optimisation are required before the induction of valuable knowledge can be automatised and taken to practical use. In order to apply the experiment on a larger scale, a more generic infrastructure is needed to support the automatic data extraction, preparation, modelling and evaluation. The input data scalability is manageable, but pattern space scalability is insufficiently dealt with.

With regard to scaling up the application, further research is needed that focuses on algorithm and representation design techniques to optimise the search. Until now, regularities were searched for in order to deduce irregularities afterwards. So a first approach to scale up is to integrate a direct search for irregularities within the process. A second opportunity is to embody the hierarchical structure of the data in the mining process. The geographic data can be represented in function of a hierarchy ranging from fine-grained up to coarse grained descriptors and the search process can be organised correspondingly. A third opportunity is to take advantage of the spatial nature of the input data by modelling it in a graph format. In our case, graph edges could for example represent object closeness or even object similarity.

Acknowledgements

This research is part of an R&D project funded by IWT (050730). I am grateful for suggestions made by Katja Verbeeck (KaHo Sint-Lieven). I wish to thank Gert Vervaet and Frank Maes (Tele Atlas) for providing data samples and for the interesting discussions that we had on the results. Special thanks to prof. dr. Luc De Raedt (K.U.Leuven) and prof. dr. Hendrik Blockeel (K.U.Leuven) for the software support and interesting suggestions.

References

1. Plantevit, M., Goutier, S., Guisnel, F., Laurent, A., Teisseire, M.: Mining unexpected multidimensional rules. In Song, I.Y., Pedersen, T.B., eds.: DOLAP. (2007) 89–96
2. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: Lof: Identifying density-based local outliers. In Chen, W., Naughton, J.F., Bernstein, P.A., eds.: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, May 16-18, 2000, Dallas, Texas, USA, ACM (2000) 93–104
3. Knorr, E.M., Ng, R.T., Tucakov, V.: Distance-based outliers: Algorithms and applications. VLDB Journal: Very Large Data Bases **8**(3–4) (2000) 237–253

4. Aggarwal, C.C., Yu, P.S.: Outlier detection for high dimensional data. In: SIGMOD Conference. (2001)
5. Koperski, K., Han, J.: Discovery of spatial association rules in geographic information databases. In Egenhofer, M.J., Herring, J.R., eds.: Proc. 4th Int. Symp. Advances in Spatial Databases, SSD. Volume 951., Springer-Verlag (1995) 47–66
6. Malerba, D., Esposito, F., Lisi, F., Appice, A.: Mining spatial association rules in census data. Research in Official Statistics **5**(1) (2002) 19–44
7. Appice, A., Ceci, M., Lanza, A., Lisi, F.A., Malerba, D.: Discovery of spatial association rules in geo-referenced census data: A relational mining approach. Intelligent Data Analysis **7** (2003) 541–566
8. Lisi, F.A., Malerba, D.: Inducing multi-level association rules from multiple relations. Machine Learning **55**(2) (2004) 175–210
9. Malerba, D., Esposito, F., Lanza, A., Lisi, F., Appice, A.: Empowering a gis with inductive learning capabilities: The case of ingens. Journal of Computers, Environment and Urban Systems **27** (2003) 265–281
10. Ceci, M., Appice, A.: Spatial associative classification: propositional vs structural approach. J. Intell. Inf. Syst. **27**(3) (2006) 191–213
11. Kuramochi, M., Karypis, G.: Finding frequent patterns in a large sparse graph. Data Min. Knowl. Discov. **11**(3) (2005) 243–271
12. Provost, F.J., Kolluri, V.: A survey of methods for scaling up inductive algorithms. Data Mining and Knowledge Discovery **3**(2) (1999) 131–169
13. Shekhar, S., Zhang, P., Huang, Y., Vatsavai, R.R.: Trends in Spatial Data Mining. In: Data Mining: Next Generation Challenges and Future Directions. AAAI/MIT Press (2003)
14. Džeroski, S.: Multi-relational data mining: an introduction. SIGKDD Explor. Newsl. **5**(1) (2003) 1–16
15. Lavrač, N., Džeroski, S.: Inductive Logic Programming: Techniques and Applications. Ellis Horwood, New York (1994)
16. De Raedt, L., Blockeel, H., Dehaspe, L., Van Laer, W.: Three companions for data mining in first order logic. In Džeroski, S., Lavrač, N., eds.: Relational Data Mining. Springer-Verlag (2001) 105–139
17. Stolle, C., Karwath, A., De Raedt, L.: Classic'cl: An integrated ilp system. In: Proceedings of the 8th International Conference of Discovery Science, Springer-Verlag (2005) 354–362
18. Blockeel, H., De Raedt, L., Jacobs, N., Demoen, B.: Scaling up inductive logic programming by learning from interpretations. Data Mining and Knowledge Discovery **3**(1) (1999) 59–93
19. Dehaspe, L.: Frequent Pattern Discovery in First-Order Logic. PhD thesis, Department of Computer Science, Katholieke Universiteit Leuven, Belgium (1998)
20. Maervoet, J.: Rule induction for geographical databases. Master's thesis, Vrije Universiteit Brussel (2007)